Jurnal Teknologi, 57 (Sciences & Engineering) Suppl 1, March 2012: 57–67

IMPUTATION OF MISSING DATA WITH DIFFERENT MISSINGNESS MECHANISM

HO MING KANG[1], FADHILAH YUSOF[2*] & ISMAIL MOHAMAD[3]

Abstract. This paper presents a study on the estimation of missing data. Data samples with different missingness mechanism namely Missing Completely At Random (MCAR), Missing At Random (MAR) and Missing Not At Random (MNAR) are simulated accordingly. Expectation maximization (EM) algorithm and mean imputation (MI) are applied to these data sets and compared and the performances are evaluated by the mean absolute error (MAE) and root mean square error (RMSE). The results showed that EM is able to estimate the missing data with minimum errors compared to mean imputation (MI) for the three missingness mechanisms. However the graphical results showed that EM failed to estimate the missing values in the missing quadrants when the situation is MNAR.

Keywords: Missing data; expectation maximization; mean imputation

Abstrak. Kertas kerja ini mempersembahkan kaedah menganggar data ketakdapatan. Data ketakdapatan yang bermekanisma Ketakdapatan Secara Rawak Sepenuhnya (MCAR), Ketakdapatan Secara Rawak (MAR) dan Ketakdapatan Secara Tak Rawak (MNAR) disimulasikan. Algoritma Expectation Maximization (EM) dan gentian min (MI) telah digunakan dalam set data ini dan dibanding dengan menggunakkan min ralat absolut (MAE) dan punca min kuasa dua ralat (RMSE). Hasil daripada kajian menunjukkan EM dapat menganggar data ketakdapatan dengan menghasilkan ralat yang rendah berbanding dengan MI bagi ketiga-tiga mekanisme. Walau bagaimanapun, keputusan dalam graf telah menunjukkan EM gagal menganggar data ketakdapatan kuadran yang tinggi dan apabila mekanisme itu ialah MNAR.

Kata kunci: Data ketakdapatan; expectation maximization; gantian min

1.0 INTRODUCTION

Missing data is a problem that always exists in the studies of hydrology and social science. The existences of missing data are normally caused by the technical errors

[1,2&3] Department of Mathematical Sciences, Faculty of Science, Universiti Teknologi Malaysia, 81310 UTM Johor Bahru, Johor Darul Ta'azim, Malaysia

* Corresponding author: fadhilahy@utm.my

that normally happened in measurement and recording, and also because of insufficient samples used. In general, traditional methods such as listwise deletion and pairwise deletion are used before the data is analyzed. However these methods may affect the reliability of the model as the analysis will become biased and loss of precision since the missing data may relate to the observed data and the deletion of missing data will cause the sample size to become smaller and thus increase the variance of the data.

Many imputation methods have been used to deal with missing data. Acuna and Rodriguez [1] evaluated four methods of imputation: case deletion, mean imputation, median imputation and k-nearest neighbor (KNN) imputation in twelve datasets from Machine Learning Database Repository. Meanwhile in [2], they compared complete case analysis with multiple imputation in dealing with missing covariate data in medical research and showed that complete case analysis can lead to biased results as compared to multiple imputation. Study in [3] has used four different aggressive methods (simple substitution, parametric and ranked regression, and Theil method) to treat the missing rainfall data in Candelaro River Basin, Italy. Besides, the application of expectation maximization (EM) to the missing precipitation series in Turkey is presented in [4] before the homogeneity tests are used. The EM was also used in [5] and compared with the combination of auto-associative neural network and genetic algorithm. The results indicated that EM performed better when there is little dependency among the variables.

In this study, EM algorithm is used in the estimation of missing data. Three missingness mechanisms are applied in this study, namely missing completely at random (MCAR), missing at random (MAR) and missing not at random (MNAR). Simulation data with MCAR, MAR and MNAR situations are used and their performances are compared with mean imputation (MI) method. Mean imputation (MI) is used because it is the simplest approach that can preserve the mean of the data and produce smaller standard error as compared to other traditional methods [6].

2.0 DATA SIMULATION WITH DIFFERENT MISSINGNESS MECHANISM

Data of difference missingness mechanisms are simulated. In definition, the data is missing completely at random (MCAR) if the probability of data missing is

unrelated to the data itself and other variables meanwhile missing at random (MAR) is referred to the probability of missing data that is dependent on some other variables but not on the missing data itself. Lastly, the missing not at random (MNAR) is a mechanism where the probability of the data missing is dependent upon the missing data itself [7]. Therefore, a dataset with MCAR, MAR and MNAR situations is generated respectively, based on the algorithm from [8] and [9], to evaluate the performance of the imputation methods used.

A complete data of X, Y and Z is simulated based on the concept of linear regression (Eq. 1 and 2). In this case, X is a random variable that has correlation with variable Y and variable Z, meanwhile Y and Z are independent to each other.

$$Y_i = \gamma_0 + \gamma_i X_i + \varepsilon_{Yi} \tag{1}$$

$$Z_i = \lambda_0 + \lambda_1 X_i + \varepsilon_{Zi} \tag{2}$$

where $\varepsilon_{Yi} \sim N(0, \sigma^2_{\varepsilon y})$, $\varepsilon_{zi} \sim N(0, \sigma^2_{\varepsilon z})$ and $X \sim N(\mu_x, \sigma^2_x)$ .The correlation between X and Y is set to be higher (0.8) because the MAR mechanism is assumed to be dependent on variable Y. Hence, the correlation between X and Z is set at a moderate value (0.5) due to this assumption. All the parameters are obtained by using least square method,

$$\gamma_1 = \rho_{xy} \frac{\sigma_y}{\sigma_x} \tag{3}$$

$$\lambda_1 = \rho_{xz} \frac{\sigma_z}{\sigma_x} \tag{4}$$

where ρ_{xy} and ρ_{xz} are the correlation between X and Y, and correlation between X and Z respectively, σ_x, σ_y and σ_z are the standard deviation of X, Y and Z respectively. Since the errors of Eqn. (1) and (2) have a constant variance, therefore the variances are estimated as

$$Var(\varepsilon_y) = \sigma^2_y (1 - \rho^2_{xy}) \tag{5}$$

$$Var(\varepsilon_z) = \sigma^2_z (1 - \rho^2_{xz}) \tag{6}$$

Using the information above, sample size of 100 for data X, Y and Z are generated based on the linear regression of Eqn. (1) and (2). Next, different missingness mechanism (MCAR, MAR or MNAR) is then imposed to the simulated complete data by using Eq. (7) – Eq. (10). In this study, X is assumed to have missing values, meanwhile Y and Z are fully observed. Hence the probability of X to be missing can be written in a logistic regression,

$$P(X \text{ be missing} \mid \text{data}) = \frac{\exp(a_1 + a_2 Y + a_3 Z + a_4 X)}{1 + \exp(a_1 + a_2 Y + a_3 Z + a_4 X)} \tag{7}$$

$$a_1 = \ln\left(\frac{p}{1-p}\right) - a_2 \overline{Y} - a_3 \overline{Z} - a_4 \overline{X} \tag{8}$$

where a_2, a_3 and a_4 are used to control the desired missingness mechanism, meanwhile a_1 is the intercept of the points and can be evaluated as a control of the proportion, p of X to be missing. The regression line is needed to shift to the left or right in order to get the desired proportion of missing data, and it is done by further generating an uniform random number, u and obtained the difference D,

$$D = \ln\left(\frac{u}{1-u}\right) - P(X \text{ be missing} \mid \text{data}) \tag{9}$$

$$P(X \text{ be missing} \mid \text{data}) = \frac{\exp(a_1 + a_2 Y + a_3 Z + a_4 X + D)}{1 + \exp(a_1 + a_2 Y + a_3 Z + a_4 X + D)} \tag{10}$$

The values from Eq. (10) and u are sorted. The highest probability is actually the desired proportion of X to be missing. Therefore the corresponding X are then made to be artificially missing to get the desired missingness mechanism. Table 1 shows the different values of a_2, a_3 and a_4 that represents the types of missingness mechanism.

Table 1 Types of missingness mechanism

Missingness Mechanism	a_2	a_3	a_4
MCAR	0	0	0
MAR on Y	0.2	0	0
Strong MAR on Y	2	0	0
MNAR on X	0	0	0.2
Strong MNAR on X	0	0	2

3.0 METHODS

In this study, expectation maximization (EM) algorithm is used to compare with the mean imputation (MI) in dealing with the missing values in X. The performances of the methods are evaluated by computing the mean absolute error (MAE) and root mean square error (RMSE).

3.1 Expectation Maximization (EM) Algorithm

The EM algorithm is a general method used to find model parameters for incomplete data. The algorithm starts with the E-step that concentrates on finding the expectation of the log-likelihood function that is conditional on the observed data and the model parameters. It is followed by the M-step where the log-likelihood function is then maximized to find the new model parameters. The algorithm will iterate until the estimated parameters converge. The likelihood of the multivariate normal (*X, Y, Z*) is given as,

$$L(\mu_x, \mu_y, \mu_z, \sigma_x^2, \sigma_y^2, \sigma_z^2) = \prod_{i=1}^{N} \frac{1}{\sigma_x^2} \exp\left\{-\left(\frac{x_i - \mu_x}{\sigma_x}\right)^2\right\} \text{x } \frac{1}{\sigma_y^2} \exp\left\{-\left(\frac{y_i - \mu_y}{\sigma_y}\right)^2\right\}$$
$$\text{x } \frac{1}{\sigma_z^2} \exp\left\{-\left(\frac{z_i - \mu_z}{\sigma_z}\right)^2\right\}$$

The log-likelihood is then become,

$$\ln\ L(\mu_x, \mu_y, \mu_z, \sigma_x^2, \sigma_y^2, \sigma_z^2, \beta_0, \beta_1) = \ln\left(\frac{1}{\sigma_x^2 \sigma_y^2 \sigma_z^2}\right)^n + \sum_{i=1}^{N}\left\{-\left(\frac{x_i - \mu_x}{\sigma_x}\right)^2\right\}$$
$$+ \sum_{i=1}^{N}\left\{-\left(\frac{y_i - \mu_y}{\sigma_y}\right)^2\right\} + \sum_{i=1}^{N}\left\{-\left(\frac{z_i - \mu_z}{\sigma_z}\right)^2\right\}$$

and since $\mu_x = \mu_x$, $\mu_y = \gamma_0 + \gamma_1 X_i$ and $\mu_z = \lambda_0 + \lambda_1 X_i$, therefore

$$\ln L(\mu_x, \mu_y, \mu_z, \sigma_x^2, \sigma_y^2, \sigma_z^2, \beta_0, \beta_1) = \ln\left(\frac{1}{\sigma_x^2 \sigma_y^2 \sigma_z^2}\right)^n - \left(\frac{1}{\sigma_x^2} + \frac{\gamma_1^2}{\sigma_y^2} + \frac{\lambda_1^2}{\sigma_z^2}\right)\sum_{i=1}^{N} x_i^2$$
$$- \frac{1}{\sigma_y^2}\sum_{i=1}^{N} y_i^2 - \frac{1}{\sigma_z^2}\sum_{i=1}^{N} z_i^2 + 2\left(\frac{\mu_x}{\sigma_x^2} + \frac{\gamma_0\gamma_1}{\sigma_y^2} + \frac{\lambda_0\lambda_1}{\sigma_z^2}\right)\sum_{i=1}^{N} x_i$$
$$+ 2\left(\frac{\gamma_0}{\sigma_y^2}\right)\sum_{i=1}^{N} y_i + 2\left(\frac{\lambda_0}{\sigma_z^2}\right)\sum_{i=1}^{N} z_i + 2\left(\frac{\gamma_1}{\sigma_y^2}\right)\sum_{i=1}^{N} x_i y_i$$
$$+ 2\left(\frac{\lambda_1}{\sigma_z^2}\right)\sum_{i=1}^{N} x_i z_i - n\left\{\left(\frac{\mu_x}{\sigma_x}\right)^2 + \left(\frac{\gamma_0}{\sigma_y}\right)^2 + \left(\frac{\lambda_0}{\sigma_z}\right)^2\right\}$$

and it is also proven that this belongs to the exponential family. In this model, E-step of EM is calculated by Eq. (11) – (14),

$$E(\sum X \mid data) = \sum_{i \in V} X_i + \sum_{i \in \overline{V}} E(X_i \mid y_i, z_i) \tag{11}$$

$$E(\sum XY \mid data) = \sum_{i \in V} X_i Y_i + \sum_{i \in \overline{V}} y_i E(X_i \mid y_i, z_i) \tag{12}$$

$$E(\sum XZ \mid data) = \sum_{i \in V} X_i Z_i + \sum_{i \in \overline{V}} z_i E(X_i \mid y_i, z_i) \tag{13}$$

$$E(\sum X^2 \mid data) = \sum_{i \in V} {X_i}^2 + \sum_{i \in \overline{V}} E({X_i}^2 \mid y, z) + Var(X \mid y_i, z_i) \tag{14}$$

In the M-step, the parameters of the model are then estimated

$$\begin{pmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \end{pmatrix} = \begin{pmatrix} s_{yy} & s_{yz} \\ s_{yz} & s_{zz} \end{pmatrix}^{-1} \begin{pmatrix} s_{xy} \\ s_{xz} \end{pmatrix} \tag{15}$$

$$s_{xx} = \frac{\left[E(\sum X^2 \mid data) - \frac{[E(\sum X \mid data)]^2}{n} \right]}{n-1} \tag{16}$$

$$s_{xy} = \frac{\left[E(\sum XY \mid data) - \frac{E(\sum X \mid data) \sum Y}{n} \right]}{n-1} \tag{17}$$

$$s_{xz} = \frac{\left[E(\sum XZ \mid data) - \frac{E(\sum X \mid data) \sum Z}{n} \right]}{n-1} \tag{18}$$

$$s_{yz} = \frac{s_{xy} s_{xz}}{s_{xx}} \tag{19}$$

$$Var\left[X_i - E(X \mid y_i, z_i) \right] \text{ for } i \in V \tag{20}$$

$$\hat{\beta}_0 = \overline{X} - \hat{\beta}_1 \overline{Y} - \hat{\beta}_2 \overline{Z} \tag{21}$$

where $\sum Y$ and $\sum Z$ are fully obtained from data, $\overline{X}$ $\overline{Y}$ and $\overline{Z}$ are the mean of X, Y and Z respectively. Finally the estimated of X is obtained as Eq. (22),

$$\hat{X} = \hat{\beta}_0 + \hat{\beta}_1 Y + \hat{\beta}_2 Z \tag{22}$$

3.2 Mean Imputation (MI)

The MI is a traditional method and the simplest method that used by hydrologists in dealing with missing data [10]. In this study, all the missing values of X will be replaced by taking the mean of the observed data, $\overline{X}$.

4.0 RESULTS AND DISCUSSION

Table 2 shows the results of mean absolute error (MAE) and root mean square error (RMSE) of 20% and 80% of missing data with MCAR, MAR and MNAR missingness. It shows that EM performs excellently as compared to mean imputation (MI) by comparing the results of MAE and RMSE. When the data is MCAR or MNAR, the errors produced by EM are comparatively higher than MAR. Therefore, EM is considered a good and excellent tool when the missingness is MAR.

Figure 1 and 2 shows the results of EM when applied to the data with strong MNAR situation. Missing data with 20% and 80% are compared. Both figures exhibit that although EM is able to estimate the missing values, the predicted values are actually underestimating the missing quadrants when compared to the complete data of X. All the highest values in missing X failed to be predicted by EM. This is also explains why the MAE and RMSE are high when MNAR is applied.

Figure 3 and 4 shows the results of EM when applied to the data with strong MAR on Y. Although the missing percentage is higher (up to 80% of missing on X), the EM is still able to predict the missing quadrants accurately. In other words, all the highest missing values is able to be estimated by EM with minimum errors. This justifies further the strength of EM in estimating the missing data when the missingness is MAR.

Table 2 Results of MAE and RMSE with different missingness mechanism on the data with 20% and 80% missing

	20% Missing				**80% Missing**			
	EM		**MI**		**EM**		**MI**	
	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE
MCAR	2.0983	2.9378	3.5242	4.2354	1.9791	2.5217	3.2129	3.9904
MAR on Y	2.0564	2.8717	3.6776	4.4586	2.0599	2.5754	3.1664	3.8632
Strong MAR on Y	1.7234	2.1269	3.9978	4.7217	1.8944	2.3807	5.3477	6.0705
MNAR on X	1.9234	2.8052	3.8706	4.6528	2.0424	2.5336	3.2548	3.9455
Strong MNAR on X	2.2997	2.9178	5.9037	6.0817	2.6761	3.2172	6.8816	7.3848

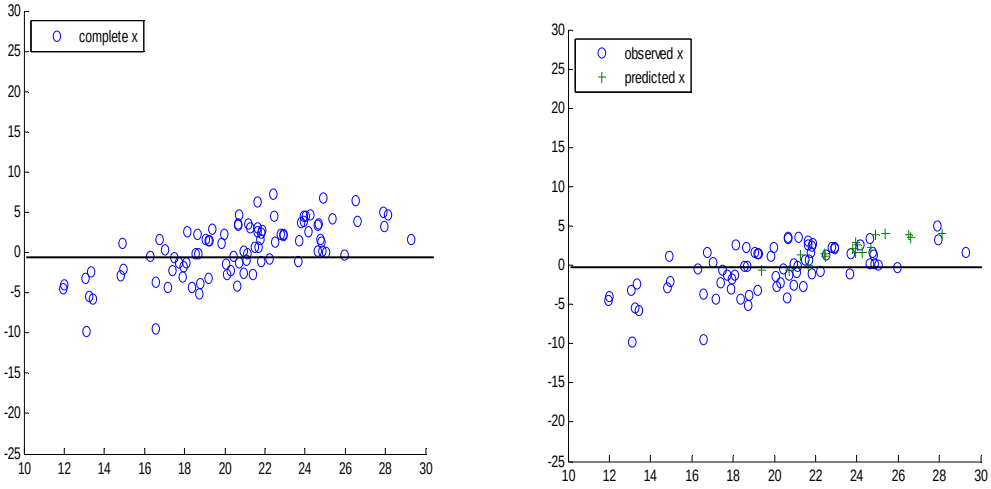

Figure 1 Comparative graph of complete and predicted missing values by EM when the data is strong MNAR with 20% missing (o is observed of *X*, + is predicted of X)

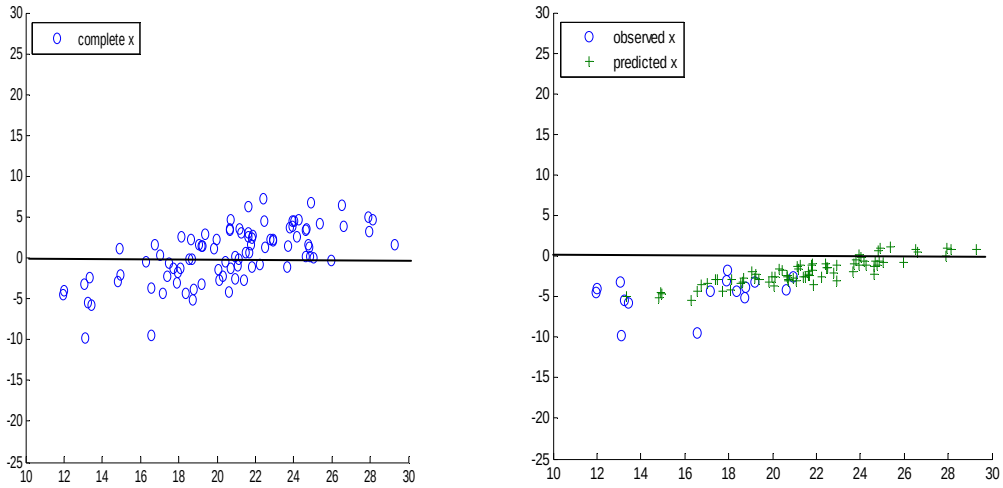

Figure 2 Comparative graph of complete and predicted missing values by EM when the data is strong MNAR with 80% missing (o is observed of *X*, + is predicted of X)

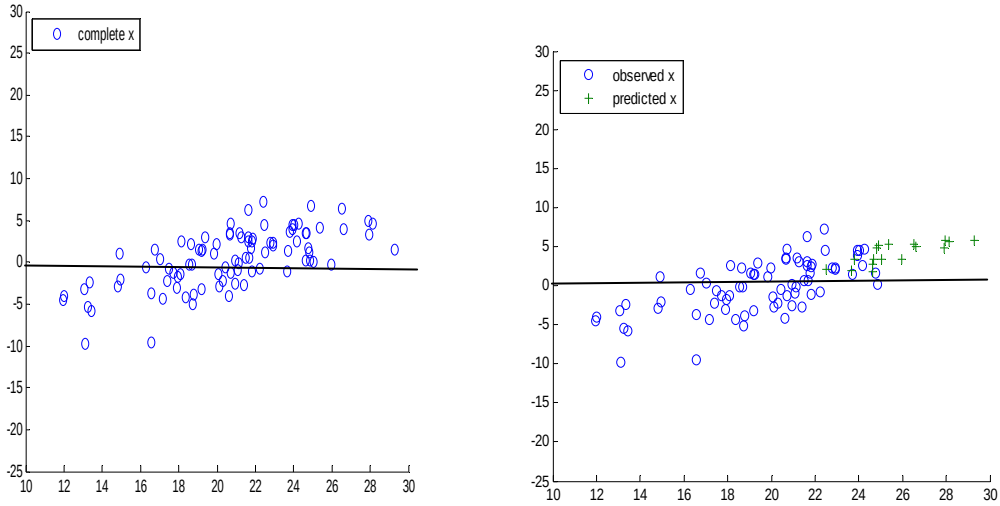

Figure 3 Comparative graph of complete and predicted missing values by EM when the data is strong MAR with 20% missing (o is observed of *X*, + is predicted of X)

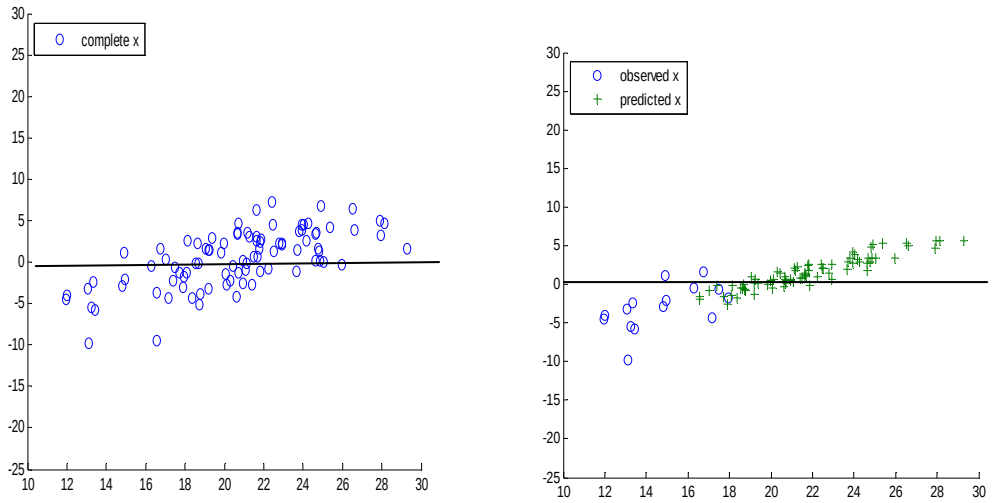

Figure 4 Comparative graph of complete and predicted missing values by EM when the data is strong MAR with 80% missing (o is observed of *X*, + is predicted of X)

5.0 CONCLUSION

This study is concerned on the comparison between EM algorithm and mean imputation (MI) on estimating missing data using simulated data with 20% and 80% of missing data. Three different missingness mechanisms namely MCAR, MAR and MNAR are artificially applied on the simulated data.

All the results indicated that EM algorithm is a powerful tool in estimating the missing values as compared to MI method. This is because the replacement of missing data by mean of observed sample will shift some high values to the middle of the distribution and thus can reduce the variance of the data. Therefore either in MCAR, MAR or MNAR, it is not encouraged to use MI since the errors produced are comparatively large.

The EM shows numerically by comparing the values of RMSE and MAE to be a good tool in predicting missing values. However the graphical results shows that EM actually failed to estimate the missing quadrants, especially when the data is having strong MNAR missingness mechanism.

As a conclusion, EM algorithm is an excellent tool when the data is missing at MCAR and MAR situation. If the data is missing at MNAR situation, EM

algorithm alone is not enough and further research is to be done to improve the estimation.

ACKNOWLEDGEMENT

The authors would like to thank the Zamalah Scholarship, Universiti Teknologi Malaysia (UTM) for the financial funding.

REFERENCES

[1] E. Acuna, C. Rodriguez. 2004. The Treatment of Missing Values and its Effect in the Classifier Accuracy. In Classification, Clustering and Data Mining Application. 639-648.

[2] M. Janssen, Donders, A. R. T., Harrell, F. E., Vergouwe, Y. *et al.* 2009. Missing Covariate Data in Medical Research: to Impute is Better than to Ignore. *Journal of Clinical Epidemiology*. 63: 721-727.

[3] R. Presti, E. Barca, G. Passarella. 2010. A Methodology for Treating Missing Data Applied to Daily Rainfall Data in the Candelaro River Basin (Italy). Environ. Monit. Assess 160: 1-22.

[4] M. Firat, F. Dikbas, A. C. Koc, M. Gungor. 2010. Missing Data Analysis and Homogeneity Test for Turkish Precipitation Series. *Sadhana*. 35(6): 707-720.

[5] F. V. Nelwamondo, S. Mohamed, T. Marwala. 2007. Missing Data: A Comparison of Neural Network and Expectation Maximization Techniques. *Current Science*. 93(11): 1514-1520.

[6] M. Nakai. 2011. Analysis of Imputation Methods for Missing Data in AR(1) Longitudinal Dataset. *Int. Journal of Math. Analysis*. 5(45): 2217-2227.

[7] R. J. A. Little, D. B. Rubin. 1987. *Statistical Analysis with Missing Data*. Unites States in America: John Wiley & Sons Inc.

[8] I. Mohamad. 2003. Data Analysis in the Presence of Missing Data. PhD Thesis, Lancaster University.

[9] W. Y. Chin, Z. M. Khalid, M. K. Ho. 2011. Analysis of Repeated Measures via Simulation. Simposium Kebangsaan Sains Matematik ke-19 (SKSM 19), UiTM Pulau Pinang, 9-11 November 2011.

[10] Y. L. Xia, P. Fabian, A. Stohl, M. Winterhalter. 1999. Forest Climatology: Estimation of Missing Values for Bavaria, Germany. *Agricultural and Forest Meteorology*. 96: 131-144.